BokehDiff: Neural Lens Blur with One-Step Diffusion

Supplementary Material

Chengxuan Zhu[1†] Qingnan Fan[2*] Qi Zhang[2] Jinwei Chen[2] Huaqi Zhang[2] Chao Xu[1] Boxin Shi[3,4*]
[1]National Key Lab of General AI, School of Intelligence Science and Technology, Peking University
[2]Vivo Mobile Communication Co., Ltd.
[3]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[4]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
{peterzhu,shiboxin}@pku.edu.cn, qingnanfan@vivo.com
https://github.com/FreeButUselessSoul/bokehdiff

A. Experimental Details

A.1. Network Design

As mentioned in Eq. (6) of the paper, we show the implementation of function $\mathrm{Soft}(\cdot)$ at the iteration of k as

$$\mathrm{Soft}(x, k) = (1 + \min(k, k_{\max}) \exp(x))^{-1}, \tag{S1}$$

where the threshold $k_{\max}$ is set as 10^6 empirically. The design rationale is to gradually approach a step function, and to keep it unchanged after the training step reaches the threshold. During inference, we fix $k = k_{\max}$.

A.2. Efficiency Comparison

As an important factor in real-world deployment, the durations of the proposed method and the baselines are compared. Profiled on EBB Val294 [2], the average interval of bokeh rendering is listed in Tab. S1. All the tests are conducted on a single NVIDIA RTX A6000 GPU, and only the duration of the model forward time is calculated. Note that MPIB [7] and Dr. Bokeh requires significantly longer time to run, because they requires per-layer inpainting in their multi-layer representations.

A.3. Datasets and Quantitative Comparisons

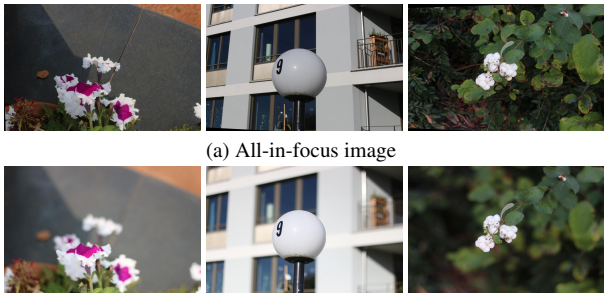
EBB Val294 is an established subset [4, 9] of the EBB! dataset [2]. It is composed of image pairs of wide and shallow depth of field captured by a DSLR camera. Though image registration is already performed [2], there are many cases where the ground truth deviates from the input in terms of global exposure level. As shown in Fig. S1, the all-in-focus image is obviously darker in the first two examples, and shows black edges near the image edge, which is caused by image registration. These artifacts combined makes the metrics of pixel-wise correspondence less persuasive in the original dataset.

For a more informed comparison, we test the performance of the images by comparing them to the original EBB Val294 [2] dataset, and the results are shown in Tab. S1. Note that as the apertures used in the EBB [2] dataset are unknown, we find the optimal aperture by binary searching, similar to the approach taken by previous methods [6, 7]. As the quantitative metrics in the paper are calculated on the exposure-aligned EBB Val294 [2] dataset, we first report the quantitative performance with the same aperture as the paper in the left columns of Tab. S1. Then we search for the optimal aperture on the original EBB Val294 [2] dataset, and list the metrics in the right columns of Tab. S1.

BLB [6], a synthetic dataset proposed by BokehMe [6], consisting of 10 scenes, and 10 focal settings for each, rendered by Blender. The rendered bokeh can be significantly larger than real-world bokeh, so it can measure the accuracy of the underlying physics model. The most challenging

Table S1. Quantitative comparison on the exposure-aligned and the original EBB Val294 [2] dataset with the same optimized aperture as the paper (left), and the aperture that is optimized in the original EBB Val294 [2] datset (right). ↑ (↓) indicates larger (smaller) values are better, and **bold** font indicates the best results. ★ denotes that the method is trained with the same dataset as BokehDiff.

Dataset		Exposure-aligned EBB Val294 [2]				Original EBB Val294 [2]			
Method	Duration (s)	PSNR↑	SSIM↑	DISTS↓	LPIPS↓	PSNR↑	SSIM↑	DISTS↓	LPIPS↓
DeepLens [3]	0.402	22.703	0.7623	0.1483	0.4191	22.065	0.7604	0.1509	0.4224
MPIB [7]	31.87	23.334	0.7920	0.1581	0.4031	22.450	0.7892	0.1616	0.4056
BokehMe [6]	1.531	24.014	0.8134	0.1460	0.3921	23.247	0.8117	0.1463	0.3918
Dr.Bokeh [8]	99.67	23.479	0.8221	0.1225	0.3771	21.298	0.8061	0.1338	0.3878
Restormer★ [10]	0.962	23.960	0.7961	0.1297	0.3778	23.188	0.7964	0.1314	0.3801
BokehMe★ [6]	1.531	23.753	0.7919	0.1437	0.3967	22.857	0.7886	0.1458	0.3998
BokehDiff	3.974	**24.652**	**0.8357**	**0.1155**	**0.3737**	**23.728**	**0.8390**	**0.1148**	**0.3711**



(a) All-in-focus image

(b) Lens blur image (real)

Figure S1. Examples from the original EBB [2] dataset that shows misalignment. In the first two example, the lens blur image has clearly more exposure than the all-in-focus image; In the second and the third example, the images show some black edges near the image border, which is caused by image registration.

Table S2. Quantitative comparison on the SYNBOKEH300 dataset. ↑ (↓) indicates larger (smaller) values are better, and **bold** font indicates the best results. ★ denotes that the method is trained with the same dataset as BokehDiff.

Method	PSNR↑	SSIM↑	DISTS↓	LPIPS↓
DeepLens [3]	24.824	0.8121	0.1403	0.3218
MPIB [7]	31.588	0.9465	0.0499	0.1129
BokehMe [6]	33.357	0.9532	0.0459	0.1129
Dr.Bokeh [8]	30.157	0.9532	0.0682	0.1504
Restormer★ [10]	32.016	0.9220	0.0695	0.1695
BokehMe★ [6]	31.329	0.9403	0.0641	0.1231
BokehDiff	**34.165**	**0.9784**	**0.0433**	**0.1119**

level 5 is used for evaluation.

SYNBOKEH300, a new synthetic benchmark generated as described in the paper. It is composed of 300 images, at 4 levels of different lens blur strengths, the ground truth disparity map, focus distance, and the all-in-focus input images. The dataset excels others in terms of photorealism and diversity, and can be used to evaluate the performance in real-world scenarios. The results are listed in Tab. S2.

B. Explanation on the EBB Dataset

Note that given a method, we select the "best" result by selecting the aperture parameters that has the best SSIM performance. As the original EBB Val294 [2] dataset is not aligned in exposure level, it may hinder the optimal exposure level selection process. As observed in Tab. S1, both of the two groups are tested on the original EBB Val294 [2] dataset, but the LPIPS [11] performance of the group optimized on exposure-aligned EBB Val294 [2] dataset is obviously superior. As discussed in the paper, LPIPS [11] is more sensitive to blurriness by design, and less sensitive to pixel-level difference. The efficacy of BokehDiff is further proved by the performance listed in Tab. S1.

C. Samples of the Data Synthesis Pipeline

Figure S2. Diverse scenes sampled from the SYNBOKEH300 dataset, to verify its photorealism and diversity.

To demonstrate the results of the proposed data synthesis, we provide some samples of the SYNBOKEH300 dataset. In Fig. S2, the scene diversity of the SYNBOKEH300 dataset is demonstrated sufficiently. In Fig. S3, we further show that the synthesis pipeline can generate both background-focused and foreground-focused images photo-realistically.

Note that the ability of BokehDiff to focus on any specific depth (as shown in the paper and later in Appendix D) originates from the training data. By randomly placing the location and facing angles, the rendered data contains progressively blur with respect to the changing depth, as well as the different amount of blur caused by the disparity offset from the focal plane.

D. More Results

D.1. Adjusting Aperture

We first demonstrate the results of increasing the blurriness in Fig. S4. In the first case, BokehDiff successfully creates the desired progressive blurriness, and in the second case, manages to blur both the foreground and the background that are off the focal plane. In both examples, BokehDiff is able to follow the underlying physics rules, and creates the right results at depth discontinuities.

D.2. Adjusting Focus Distance

We provide another example of changing focus distance in Fig. S5. As the error is more subtle when the background and foreground are both out of focus, we mainly present the images that focuses on the foreground.

(a) Disparity

(b) All-in-focus

(c) Synthetic ground truth with different aperture parameters and different focus distance.

Figure S3. An example of the synthetic data. With the mechanism described in the paper, we can get (a) disparity map, (b) all-in-focus image as input, and (c) synthetic ground truth images under different apertures and focus distance settings. Note that the first row in (c) is focused on the background, and the second is on the foreground, with aperture growing larger from left to right.

(a) Inputs (b) 1x blur (c) 2x blur (d) 3x blur

Figure S4. Given the defocus map and all-in-focus image shown in (a), we demonstrate the results of gradually increasing the aperture parameters, from (b) 1x blurriness to (c) 3x blurriness. Please zoom in for details.

D.3. Comparisons

Here we provide some more comparisons of BokehDiff and the baselines, to further validate the efficacy of BokehDiff.

First we demonstrate some more comparisons in Fig. S6. In the first example, BokehDiff successfully focuses on the person, and creates a progressive blurrness for the ground behind and before the person. In comparison, BokehMe [6] over-blurs the person's helmet and hands due to the erroneous disparity estimation. MPIB [7] fails to produce the progressive blur, while Dr. Bokeh [8] creates unnatural split near the boundary of the person.

In the second example, all the baselines over-blurs the thin end of the man's beard, while BokehDiff keeps the focused detail intact. As for the blurry foreground, BokehDiff creates a physically correct and beautiful semi-transparent blur near the unfocused edge of the sleeves. In comparison, the baselines either create a hard edge (BokehMe [6] and Dr. Bokeh [8]) or over-blur the boundary (MPIB [7]).

The third example is another case where BokehDiff outperforms previous methods at depth discontinuities. Most

Figure S5. A synthetic focal stack of BokehDiff, given an all-in-focus image selected from the Unsplash [1] dataset.

of the hair of the person should be focused, but the baselines over-blur the part near the edge, while BokehDiff manages to keep the fine details in focus. The transition to out-of-focus area is also smooth and natural.

We continue the demonstration of results in Fig. S7. In the first example BokehDiff keeps the thin details of the cat's fur while blurring the window behind them, while the other methods show different degrees of artifacts. In the second example, BokehDiff manages to create a progressive blur as the defocus increases, while keeping the focused foreground intact, even in such area as the hair seam and the elbow where the background is messy. In comparison, the baselines follow the inaccurate depth estimation result, and create bumps near the hair, and unnatural zig-zags near the elbow. The hair in the green box is also blurred by mistake.

The third example also demonstrates the effectiveness of BokehDiff in generating realisitic lens blur for intricate structures. The progressive blur in the background also validates that BokehDiff follows the image formation model.

In Fig. S8, as shown from the first two examples, the proposed method also work on images shot with a wide aperture, and further blurs the blurred background while keeping the foreground in focus. Both the hair streaks of the person and the furs of the cat are effectively kept. In the third example, BokehDiff also show the ability to synthesize progressive blurriness, while keeping the person's beard and the focused T-shirt. In comparison, the baselines cannot preserve the intricate details, as well as the regions where depth estimation methods go wrong, such as the hair of the person in the first example, the fur near the cat's ear in the second example, and the T-shirt edge in the third example.

To sum up, given all the demonstrated results, we conclude that the results rendered by BokehDiff are both physically reasonable and visually pleasant.

E. Visual Results of Ablation Study

We present some visual results to further support the ablation study in Fig. S9.

As the PISA module is designed to bring in constraints related to physics, an incomplete version cannot model the image formation model by design, and thus can only resort to learning from the data distribution.

We first try removing the circle-of-confusion constraint, and the result in Fig. S9(b) looks very similar to the all-in-focus input, indicating that the PISA module determines how much the blurriness should be.

In Fig. S9(c), removing the self-occlusion from the PISA module blurs the foreground that should be in focus, which is caused by the ignorance of keeping the foreground before the background, since self-occlusion is removed.

As for the energy-conserved normalization, since the energy no longer follows the physics intuition, the results look blurry even for the focused region in the green box of Fig. S9(d).

F. Future Works

We conclude the paper with the some ideas for thought, hoping that BokehDiff inspire more interesting works.

The realm of lens blur rendering still lacks a metric to measure the "photorealism". With paired data, LPIPS is found to be the most sensitive to wrong blurring pattern [11]. However, it is easy for human vision system to see the lens blur is synthesized, even without ground truth. The prior behind such phenomenon is intriguing, and requires further analysis.

For example, will it be possible to train a discriminator that is able to focus on the low-level details that renders the image to be "fake" to human eyes, as an important reference-free metric? If so, can we iterate a generator over the discriminator, to yield even more realistic images?

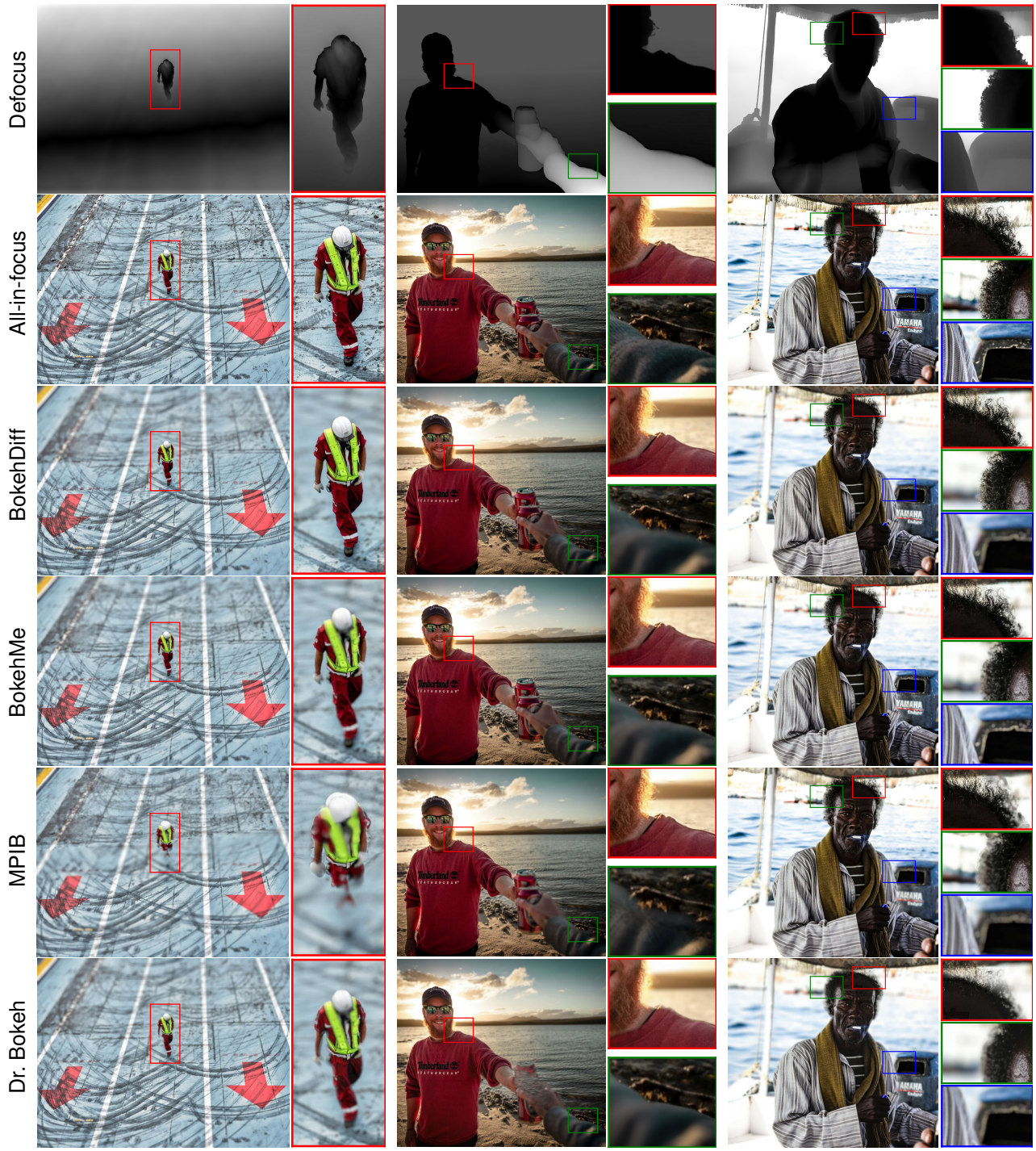


Figure S6. More qualitative comparisons of BokehDiff with BokehMe [6], MPIB [7], and Dr. Bokeh [8]. Calculated from disparity, the defocus map is shared across the methods to be compared. The defocus map is for reference only, with whiter regions for more lens blur, but is subjected to error caused by depth estimation.

In addition, photorealistic video lens blur rendering is also an interesting follow-up thread, with its unique challenges such as consistency. With the proposed data synthesis pipeline, it will be easier to train a similar video bokeh rendering method, but this idea is beyond the scope of the paper, and deserves a paper of its own.

As for the model design, the PISA module requires more investigation. The biggest difference is that it changes the dimension on which to perform normalization. It is not self-evident to scale the default normalization to larger models (such as DiT [5]).

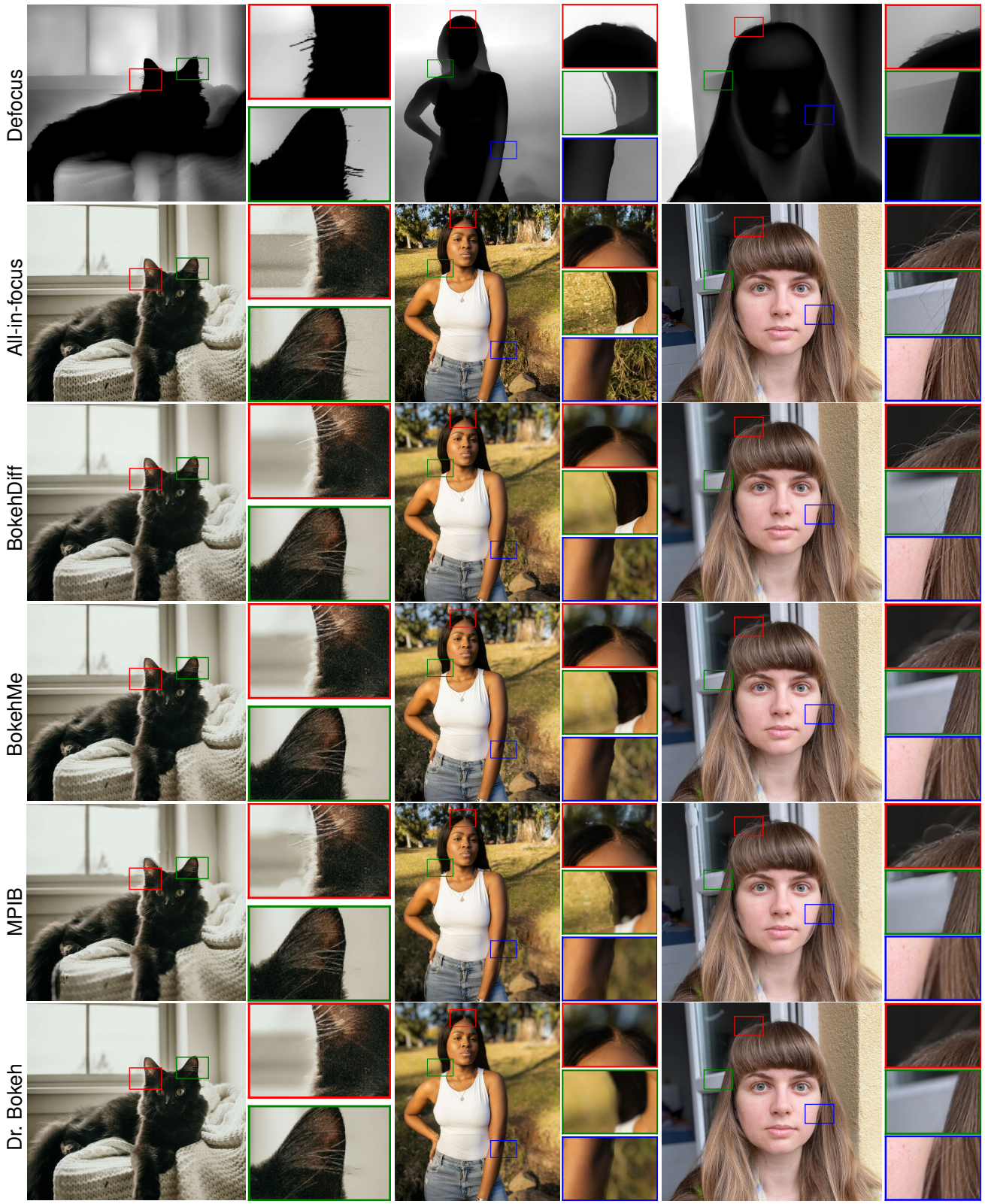


Figure S7. More qualitative comparisons of BokehDiff with BokehMe [6], MPIB [7], and Dr. Bokeh [8]. Calculated from disparity, the defocus map is shared across the methods to be compared. The defocus map is for reference only, with whiter regions for more lens blur, but is subjected to error caused by depth estimation.

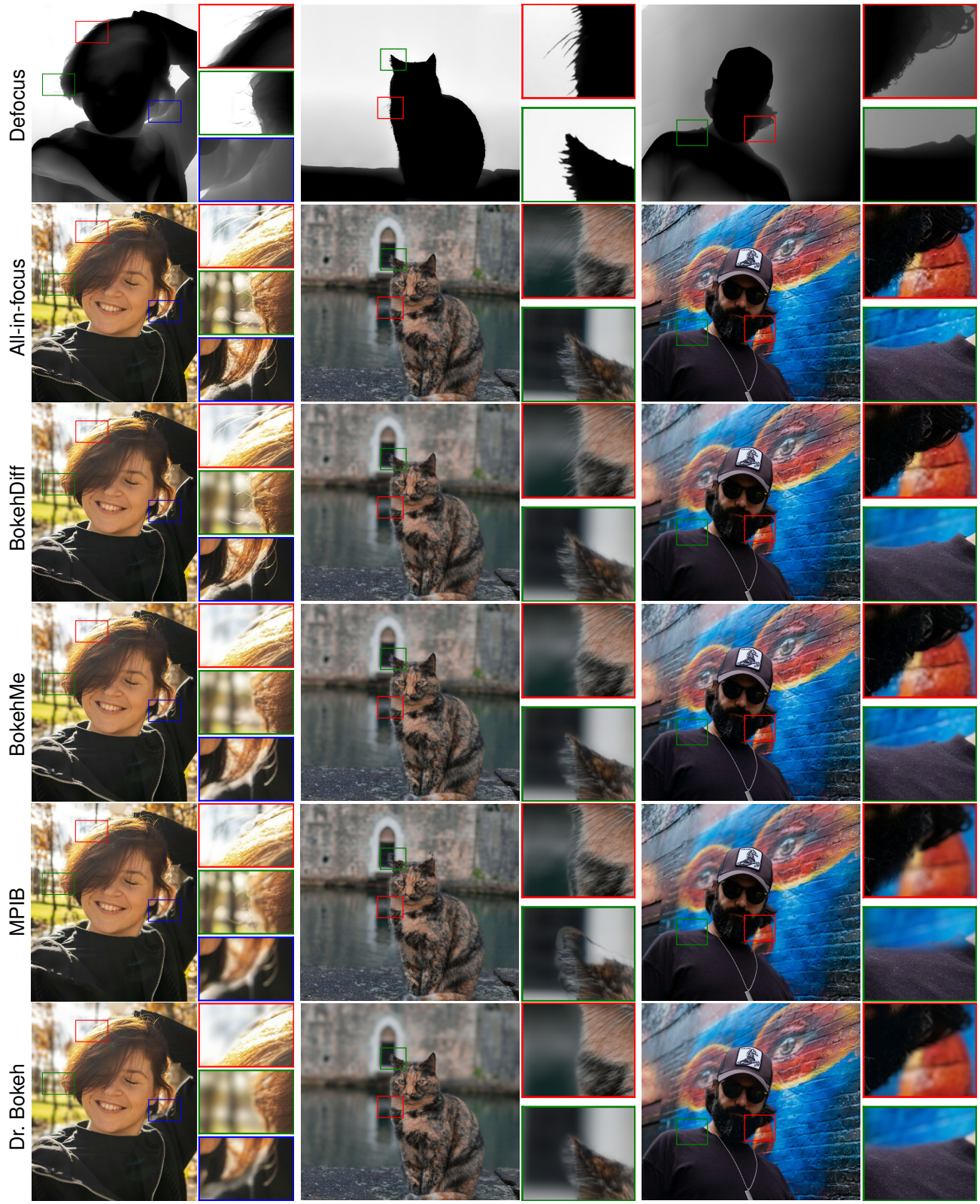


Figure S8. More qualitative comparisons of BokehDiff with BokehMe [6], MPIB [7], and Dr. Bokeh [8]. Calculated from disparity, the defocus map is shared across the methods to be compared. The defocus map is for reference only, with whiter regions for more lens blur, but is subjected to error caused by depth estimation.

(a) All-in-focus (b) W/o CoC (c) W/o occlusion (d) W/o SoftmaxQ (e) Complete Model

Figure S9. Visual comparisons of the ablation study. The setting of "SoftmaxQ", "CoC", and "occlusion" are short for the energy-conserved normalization, circle of confusion constraint, and self-occlusion respectively.

References

[1] Luke Chesser, Timothy Carbone, and Ali Zahid. Unsplash full dataset 1.2.2, 2020. unsplash.com/data, Last accessed on 2024-10-31. 4

[2] Andrey Ignatov, Radu Timofte, Ming Qian, Congyu Qiao, Jiamin Lin, Zhenyu Guo, Chenghua Li, Cong Leng, Jian Cheng, Juewen Peng, et al. Aim 2020 challenge on rendering realistic bokeh. In *Proc. of European Conference on Computer Vision Workshops*, 2020. 1, 2

[3] Wang Lijun, Shen Xiaohui, Zhang Jianming, Wang Oliver, Lin Zhe, Hsieh Chih-Yao, Kong Sarah, and Lu Huchuan. DeepLens: Shallow depth of field from a single image. *ACM Transactions on Graphics*, 37(6), 2018. 1, 2

[4] David Mandl, Shohei Mori, Peter Mohr, Yifan Peng, Tobias Langlotz, Dieter Schmalstieg, and Denis Kalkofen. Neural bokeh: Learning lens blur for computational videography and out-of-focus mixed reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces*, 2024. 1

[5] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. of International Conference on Computer Vision*, 2023. 5

[6] Juewen Peng, Zhiguo Cao, Xianrui Luo, Hao Lu, Ke Xian, and Jianming Zhang. BokehMe: When neural rendering meets classical rendering. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 5, 6, 7

[7] Juewen Peng, Jianming Zhang, Xianrui Luo, Hao Lu, Ke Xian, and Zhiguo Cao. MPIB: An MPI-based bokeh rendering framework for realistic partial occlusion effects. In *Proc. of European Conference on Computer Vision*, 2022. 1, 2, 3, 5, 6, 7

[8] Yichen Sheng, Zixun Yu, Lu Ling, Zhiwen Cao, Xuaner Zhang, Xin Lu, Ke Xian, Haiting Lin, and Bedrich Benes. Dr. Bokeh: Differentiable occlusion-aware bokeh rendering. In *Proc. of Computer Vision and Pattern Recognition*, 2024. 1, 2, 3, 5, 6, 7

[9] Zhihao Yang, Wenyi Lian, and Siyuan Lai. BokehOrNot: Transforming bokeh effect with image transformer and lens metadata embedding. In *Proc. of Computer Vision and Pattern Recognition*, 2023. 1

[10] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 1, 2

[11] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2, 4